Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso

Shejuty Devnath

December 13, 2024

Abstract

This project focuses on exploring and replicating the findings from the paper "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso" by Hansheng Wang, Guodong Li, and Guohua Jiang (2007). The LAD-LASSO method combines the robustness of least absolute deviation (LAD) regression with the variable selection capabilities of LASSO. According to the original paper, LAD-Lasso performs well in datasets with outliers or heavy-tailed errors, also accurately identifies important predictors. To evaluate that claim, the assessment includes a comparison with other robust shrinkage methods, namely SCAD, MCP, and LASSO. Also, with estimated tuning parameters, the LAD-lasso estimator enjoys the same asymptotic efficiency as the unpenalized LAD estimator obtained under the true model (i.e., the oracle property). Comprehensive simulation studies highlight the finite-sample performance of LAD-LASSO.

1 Introduction

Datasets with heavy-tailed errors or outliers are common in many real-world applications. This can significantly affect the reliability of statistical models. These issues often arise in the response variables, where traditional methods like ordinary least squares (OLS) may not provide accurate estimates. In such cases, the least absolute deviation (LAD) estimator offers a robust alternative. A key advantage of the LAD estimator is $\sqrt{n}$-consistency and asymptotic normality can be established without imposing moment conditions on the residuals. However, the LAD criterion involves a nonsmooth objective function, we cannot apply Taylor expansions to study its asymptotic properties. Over the years, significant progress has been made in establishing the theory of $\sqrt{n}$-consistency and asymptotic normality of LAD estimators (Bassett and Koenker (1978), Pollard (1991), Bloomfield and Steiger (1983), Knight (1998), Peng and Yao (2003), and Ling (2005). Despite these advantages, the development of robust methods for model selection is left to study.

Model selection plays an important role in regression analysis, as it directly impacts the accuracy and reliability of parameter estimates and predictions. Omitting an important explanatory variable can lead to severely biased estimates and poor predictions, while including unnecessary predictors can reduce estimation efficiency and degrade prediction accuracy. Striking the right balance is essential, making model selection a key challenge in both theoretical and practical settings (Shao 1997; Hurvich and Tsai 1989; Shi and Tsai 2002, 2004).

There is no universal agreement on the most effective model selection criteria, preferences often depend on specific applications and assumptions (Shi and Tsai 2002). However LAD-lasso method offers an advantage by consistent model selection category, it assumes that the true model is finite in dimension and is included within the set of candidate models. Under this assumption, the LAD-lasso method reliably identifies the true model (Shao (1997), McQuarrie and Tsai (1998), and Shi and Tsai (2002)). LAD-based variable selection methods are robust and effective, but they come with computational complexity. The major challenge lies in the exponential growth of potential candidate models as the number of regression variables increases. When the number of predictors is large, evaluating all possible combinations to identify the best subset becomes computationally intensive and often impractical. This limitation makes it challenging to apply these methods in high-dimensional settings without efficient optimization techniques or approximations.

To overcome the limitations of traditional model selection methods, Tibshirani (1996) introduced the least absolute shrinkage and selection operator (lasso). Lasso offers a powerful approach for simultaneously selecting important explanatory variables and estimating regression parameters. (Knight and Fu (2000), Fan and Li (2001), and Tibshirani et al. 2005). However, lasso's performance can deteriorate in the presence of heavy-tailed errors or outliers. We are putting that claim to test.

The Smoothly Clipped Absolute Deviation (SCAD) penalty, introduced by Fan and Li (2001), is a powerful regularization method in statistical modeling for variable selection and shrinkage. SCAD is designed to retain the sparsity-inducing property of Lasso, also reducing bias and providing model selection consistency. Its penalty function is non-convex and smooth, which allows it to shrink small coefficients toward zero for variable selection while leaving larger coefficients nearly unbiased. This property makes SCAD suitable for high-dimensional data.

The Minimax Concave Penalty (MCP) Zhang (2010), is a regularization method designed to improve variable selection and reduce bias in regression models. Unlike Lasso, which penalizes all coefficients equally and can shrink large coefficients too much, MCP uses a non-convex penalty that applies less shrinkage to larger coefficients. This helps maintain sparsity for small coefficients while keeping larger ones more accurate.

In this project, we aim to compare the performance of four popular regularization methods—SCAD, MCP, Lasso, and LAD-Lasso—under conditions of heavy-tailed errors and outliers. These methods are widely used for variable selection and parameter estimation, but their effectiveness can vary significantly in the presence of outliers or non-normal error distributions. By evaluating their performance across various scenarios, we seek to identify the most robust and reliable method for handling heavy-tailed outliers in regression models.

2 Absolute Shrinkage and Selection

2.1 Lasso, Lasso*, and LAD-Lasso

Consider a linear regression model:

$$y_i = x_i^\top \beta + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_i = (x_{i1}, \ldots, x_{ip})^\top$ is a p-dimensional vector of covariates, $\beta = (\beta_1, \ldots, \beta_p)^\top$ represents the regression coefficients, and ϵ_i are independent and identically distributed

(i.i.d.) random errors with a median of 0. Assume further that $\beta_j \neq 0$ for $j \leq p_0$ and $\beta_j = 0$ for $j > p_0$ for some $p_0 > 0$. This implies that the correct model includes p_0 significant variables, while the remaining $(p - p_0)$ variables are insignificant.

Traditionally, the unknown parameters in this model are estimated by minimizing the Ordinary Least Squares (OLS) criterion:

$$\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2.$$

However, to simultaneously perform variable selection and shrink unnecessary coefficients to zero, Tibshirani (1996) proposed the Lasso method, which modifies the OLS criterion by adding an L_1-penalty:

$$\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda > 0$ controls the amount of shrinkage applied to the coefficients. This approach encourages sparsity in β, effectively selecting the most relevant predictors.

Lasso uses the same tuning parameters for all the regression coefficients, so we get bias estimators (Fan and Li 2001). So we come up with modified lasso criterion as follows:

$$\text{lasso}^* = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

which allows for different tuning coefficients, making it more effective than Lasso. However, OLS criteron in LASSO* might not yield good estimates in presence of outliers. So we need further modification on LAD-Lasso criterion:

$$\text{LAD-lasso} = Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}| + n \sum_{j=1}^{p} \lambda_j |\beta_j|.$$

As can be seen, the LAD-lasso criterion, with its combination of the Least Absolute Deviation (LAD) loss and the Lasso penalty, is expected to be robust against outliers and promote sparsity in the model.

The pseudo data is following:

$$Y^{\text{new}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{p\times 1} \qquad X^{\text{new}} = \begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{bmatrix}_{p\times p}$$

where,

$$\lambda\|\beta\|_1 = \left\| \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{p\times 1} - \lambda I_p \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \right\|_1$$

The augmented data becomes

$$Y^* = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(n+p)\times 1} \qquad X^* = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \\ \lambda & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & \lambda \end{bmatrix}_{(n+p)\times p}$$

Specifically, we can construct $\{(y_i^*, x_i^*)\}$ with $i = 1, \ldots, n+p$, where $1 \leq i \leq n$, $(y_{i+j}^*, x_{i+j}^*) = (0, n\lambda_j e_j)$ for $1 \leq j \leq p$, and e_j is a p-dimensional vector with the j-th component equal to 1 and all others equal to 0.

It can be readily shown that the n+p LAD-lasso objective function, Q(β), can be expressed as:

$$\text{LAD-lasso} = Q(\boldsymbol{\beta}) = \sum_{i=1}^{n+p} |y_i^* - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}|.$$

2.2 SCAD and MCP

The smoothly clipped absolute deviation (SCAD) penalty, (Fan and Li 2001), was designed to encourage sparse solutions to the least squares problem. It is a non-convex penalty function that combines the advantages of L1 and L2 penalties. It encourages sparsity like L1 but mitigates its bias for large coefficients.

A large class of variable selection models can be described under the family of models called "penalized least squares". The general form of these objective functions is:

$$Q(\beta) = \min_{\beta} \frac{1}{2}||y - X\beta||_2^2 + \sum_{j=1}^{p} p_{\lambda,\gamma}(|\beta_j|)$$

$$p_{\lambda,\gamma}(|\beta|) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta| \leq \lambda \\ -\frac{(\beta_j^2 - 2\gamma\lambda|\beta_j| + \lambda^2)}{2(\gamma-1)} & \text{if } \lambda < |\beta_j| \leq \gamma\lambda \\ \frac{(\gamma+1)\lambda^2}{2} & \text{if } |\beta_j| > \gamma\lambda \end{cases}$$

We observe extra tuning parameter $\gamma > 2$. The SCAD penalty is indeed often defined in terms of its first derivative, p'(β). This derivative is given by:

$$p'_{\lambda}(|\beta_j|) = \begin{cases} \lambda, & \text{if } 0 < |\beta_j| \leq \lambda \\ \frac{\lambda\gamma - \beta_j}{(\gamma-1)}, & \text{if } \lambda < |\beta_j| \leq \gamma\lambda \\ 0, & \text{if } |\beta_j| > \gamma\lambda \end{cases}$$

Notice that for large values of β (where $|\beta| > a\lambda$) the penalty is constant with respect to β. In other words, after β becomes large enough, higher values of β aren't penalized more. This stands in contrast to the LASSO penalty, which has a monotonically increasing penalty with respect to $|\beta|$, this means that for large coefficient values, their LASSO estimates will be biased downwards.

The idea behind the minimax concave penalty (MCP) is very similar:

$$P_\gamma(x;\lambda) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma}, & \text{if } |x| \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |x| > \gamma\lambda, \end{cases}$$

Its derivative is

$$P'_\gamma(x;\lambda) = \begin{cases} \lambda - \frac{|x|}{\gamma} \cdot \text{sgn}(x), & \text{if } |x| \le \gamma\lambda, \\ 0, & \text{if } |x| > \gamma\lambda, \end{cases}$$

As with SCAD, MCP starts out by applying the same rate of penalization as the lasso, then smoothly relaxes the rate down to zero as the absolute value of the coefficient increases. In comparison to SCAD, however, the MCP relaxes the penalization rate immediately while with SCAD the rate remains flat for a while before decreasing.

2.3 Theoretical Properties of LAD-Lasso

To simplify the analysis, we decompose the regression coefficient as $\boldsymbol{\beta} = (\boldsymbol{\beta}'_a, \boldsymbol{\beta}'_b)'$, where $\boldsymbol{\beta}_a = (\beta_1, \ldots, \beta_{p_0})'$ and $\boldsymbol{\beta}_b = (\beta_{p_0+1}, \ldots, \beta_p)'$. The associated LAD-lasso estimator is represented as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'_a, \hat{\boldsymbol{\beta}}'_b)'$, and the corresponding LAD-lasso objective function is expressed as $Q(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$. Furthermore, we also split the covariate $\mathbf{x}_i$ into two components as $\mathbf{x}_i = (\mathbf{x}'_{ia}, \mathbf{x}'_{ib})'$, where $\mathbf{x}_{ia} = (x_{i1}, \ldots, x_{ip_0})'$ and $\mathbf{x}_{ib} = (x_{i(p_0+1)}, \ldots, x_{ip})'$. To establish the theoretical foundations of LAD-lasso, the following technical assumptions are required (Pollard 1991 ; Bloomfield and Steiger 1983; Knight 1998):

Assumption A. The error e_i has a continuous and positive density at the origin.

Assumption B. The matrix $\text{cov}(\mathbf{x}_i) = \Sigma$ exists and is positive definite.

Lemma 1 [Root-n Consistency] Suppose that $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are i.i.d. and that the linear regression model (1) satisfies Assumptions A and B. If $\sqrt{n}a_n \to 0$, then the LAD-LASSO estimator is root-n consistent. It implies that if the tuning parameters associated with the significant variables converge to 0 at a speed faster than $n^-0.5$, then

Lemma 2 [Sparsity] Under the same assumptions as Lemma 1 and the further assumption that $\sqrt{n}b_n \to \infty$, with probability tending to 1, the LAD-LASSO estimator $\hat{\beta}' = (\hat{\beta}_a, \hat{\beta}_b)'$ must satisfy $\hat{\beta}_b = 0$. Lemma 2 shows that LAD-LASSO has the ability to consistently produce sparse solutions for insignificant regression coefficients; hence, variable selection and parameter estimation can be accomplished simultaneously.

Theorem (Oracle Property): Suppose that $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, are i.i.d. observations, and the linear regression model (1) satisfies Assumptions A and B. Additionally, if $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, then the LAD-lasso estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'_a, \hat{\boldsymbol{\beta}}'_b)'$ satisfies the following properties:

1. $P(\hat{\boldsymbol{\beta}}_b = 0) \to 1$, and 2. $\sqrt{n}(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a) \to \mathcal{N}(0, 0.25\Sigma_0^{-1} f^{-2}(0))$,

where $\Sigma_0 = \text{cov}(\mathbf{x}_{ia})$ and $f(t)$ is the density of the error term e_i at $t = 0$.

Under the conditions $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, this theorem establishes that the LAD-lasso estimator is robust to heavy-tailed errors because the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_a$ is achieved without requiring moment conditions on the regression error e_i. Furthermore, the theorem demonstrates that the resulting LAD-lasso estimator shares the same asymptotic distribution as the LAD estimator derived under the true model, thus confirming the oracle property of the LAD-lasso estimator. Additionally, due to the convexity and piecewise linearity of the criterion function $Q(\boldsymbol{\beta})$, the LAD-lasso estimator properties discussed in this article are global rather than local (Fan and Li 2001).

2.4 Tuning Parameter Estimation

K-fold cross-validation (Craven and Wahba 1979; Tibshirani 1996; Fan Li 2001) is used to select the optimal tuning parameter for each model. Cross-validation divides the dataset into K equal-sized folds, then each subset serves as a validation set once while the remaining k-1 folds are used for training. The average validation error across all folds, known as the CV error, is then used to assess the performance of the model for a given tuning parameter λ. The optimal λ^* is selected as the value that minimizes the CV error.

$$\text{CV Error}(\lambda) = \frac{1}{K}\sum_{k=1}^{K} \text{Error}^{(k)}_{\text{validation}}(\lambda),$$

$$\lambda^* = \arg\min_{\lambda} \text{CV Error}(\lambda).$$

3 Simulation Results

This section summarizes the results of simulation studies that evaluated the performance of LAD-lasso with varying sample sizes and heavy-tailed errors. For comparison, we included SCAD, MCP, LASSO, and the Oracle estimator is also evaluated. In each case, cross-validation was used to determine the optimal tuning parameter. Specifically, we set $p = 8$ and $\beta = (0.5, 1.0, 1.5, 2.0, 0, 0, 0, 0)$, indicating that the first $p = 4$ regression variables are significant, while the remaining are not. The covariates X_i were generated from a standard eight-dimensional multivariate normal distribution, and the sample sizes considered were $n = 50$, $n = 100$, and $n = 200$.

The response variables were generated using the model:

$$y_i = x_i'\beta + \sigma\epsilon_i,$$

where ϵ_i was drawn from heavy-tailed distributions. We examined three types of distributions: the standard double exponential, the t-distribution with 5 degrees of freedom (t_5), and the t-distribution with 3 degrees of freedom (t_3). Two values of σ, 0.5 and 1.0, were tested to represent high and low signal-to-noise ratios.

For each parameter setting, we performed 1,000 simulation iterations to evaluate the finite-sample performance of selected models. The results are summarized in Tables 1-3, which report the proportion of correctly estimated, overestimated, and underestimated regression models, along with the average number of correctly and mistakenly estimated zeros in the same manners as done by Tibshirani (1996) and Fan and Li (2001). Additionally, the tables include the mean and median of the Mean Absolute Prediction Error (MAPE). We used *glmnet* package to produce Lasso estimate, *ncvreg* for SCAD & MCP and *regnet* package to get Lad-lasso estimates.

3.1 Analysis of the Results

From table 1 we observe that for n=50, sigma=1 LAD-LASSO has the least proportion of Underfitted, MCP has the largest proportion of correctly fitted and least overfitted model, also has the most correct number of zeros. Here, LASSO has the lowest average MAPE, SCAD has the lowest median MAPE. By looking at the results, we see that LAD-LASSO has the highest average and median MAPE. For n=100 LASSO has the least proportion of Underfitted, MCP has the largest proportion of correctly fitted and

Table 1: Simulation results for double-exponential error

σ	n	**Method**	**Underfitted**	**Correctly Fitted**	**Overfitted**	**Correct_0s**	**Incorrect_0s**	**Avg_MAPE**	**Median_MAPE**
1	50	SCAD	0.03	0.37	0.60	2.73	1.27	0.9864278	0.9813554
		MCP	0.04	0.48	0.48	3.08	0.92	0.9858442	0.9858835
		LASSO	0.01	0.08	0.91	1.62	2.38	0.9849375	0.9934226
		LAD-LASSO	0.01	0.12	0.87	1.71	2.29	1.0042576	1.0106411
		ORACLE	0.00	1.00	0.00	4.00	0.00	1.0065664	1.0147417
1	100	SCAD	0.144	0.224	0.632	2.596	1.404	0.9738753	0.9657626
		MCP	0.224	0.331	0.445	3.001	0.999	0.9762597	0.9686922
		LASSO	0.067	0.084	0.849	1.736	2.264	0.9726954	0.9638466
		LAD-LASSO	0.073	0.107	0.820	1.868	2.132	1.0173927	1.0082627
		ORACLE	0.043	0.957	0.000	4.000	0.000	1.0230228	1.0178607
1	200	SCAD	0.000	0.438	0.562	2.971	1.029	0.9958897	0.9928761
		MCP	0.003	0.644	0.353	3.317	0.683	0.9965292	0.9927605
		LASSO	0.000	0.075	0.925	1.657	2.343	0.9972112	0.9930692
		LAD-LASSO	0.000	0.075	0.925	1.664	2.336	1.0061970	1.0062798
		ORACLE	0.000	1.000	0.000	4.000	0.000	1.0070598	1.0026163
0.5	50	SCAD	0.000	0.756	0.244	3.497	0.503	0.2412762	0.2391900
		MCP	0.000	0.792	0.208	3.538	0.462	0.2413754	0.2392739
		LASSO	0.000	0.068	0.932	1.597	2.403	0.2407980	0.2377424
		LAD-LASSO	0.000	0.279	0.721	2.168	1.832	0.3346031	0.3270149
		ORACLE	0.001	0.999	0.000	4.000	0.000	0.3374376	0.3277341
0.5	100	SCAD	0.000	0.771	0.229	3.525	0.475	0.2425001	0.2411104
		MCP	0.000	0.801	0.199	3.606	0.394	0.2425660	0.2409623
		LASSO	0.000	0.065	0.935	1.615	2.385	0.2419645	0.2395428
		LAD-LASSO	0.000	0.294	0.706	2.237	1.763	0.3393158	0.3310485
		ORACLE	0.000	1.000	0.000	4.000	0.000	0.3387207	0.3311048
0.5	200	SCAD	0.000	0.775	0.225	3.538	0.462	0.2481385	0.2469781
		MCP	0.000	0.791	0.209	3.603	0.397	0.2481364	0.2469388
		LASSO	0.000	0.080	0.920	1.671	2.329	0.2483467	0.2474790
		LAD-LASSO	0.000	0.174	0.826	1.953	2.047	0.2753238	0.2733824
		ORACLE	0.000	1.000	0.000	4.000	0.000	0.2754032	0.2741967

Table 2: Simulation results for t5 error

σ	n	**Method**	**Underfitted**	**Correctly Fitted**	**Overfitted**	**Correct_0s**	**Incorrect_0s**	**Avg_MAPE**	**Median_MAPE**
1	50	SCAD	0.100	0.259	0.641	2.662	1.338	0.8986494	0.8899770
		MCP	0.161	0.362	0.477	3.015	0.985	0.9005127	0.8909149
		LASSO	0.039	0.062	0.899	1.675	2.325	0.8965090	0.8847344
		LAD-LASSO	0.050	0.100	0.850	1.858	2.142	0.9451671	0.9372972
		ORACLE	0.025	0.975	0.000	4.000	0.000	0.9565173	0.9484120
1	100	SCAD	0.032	0.319	0.649	0.763	1.237	0.924302	0.9199855
		MCP	0.043	0.520	0.437	0.108	0.892	0.9249109	0.9213792
		LASSO	0.004	0.077	0.919	0.620	2.380	0.9229038	0.9208076
		LAD-LASSO	0.006	0.091	0.903	0.712	2.288	0.9445341	0.9397998
		ORACLE	0.002	0.998	0.000	4.000	0.000	0.9501238	0.9467081
1	200	SCAD	0.003	0.456	0.541	3.056	0.944	0.9391955	0.9375900
		MCP	0.003	0.650	0.347	3.343	0.657	0.9397332	0.9390577
		LASSO	0.002	0.088	0.910	1.686	2.314	0.9397835	0.9391818
		LAD-LASSO	0.002	0.098	0.900	1.691	2.309	0.9498914	0.9462594
		ORACLE	0.001	0.999	0.000	4.000	0.000	0.9517892	0.9502863
0.5	50	SCAD	0.004	0.450	0.546	3.023	0.977	0.4492681	0.4429403
		MCP	0.009	0.635	0.356	3.304	0.696	0.4501220	0.4440357
		LASSO	0.000	0.069	0.931	1.501	2.499	0.4477034	0.4393192
		LAD-LASSO	0.001	0.160	0.839	1.878	2.122	0.5155815	0.5105417
		ORACLE	0.001	0.999	0.000	4.000	0.000	0.5218440	0.5126820
0.5	100	SCAD	0.000	0.670	0.330	3.345	0.655	0.4625746	0.4608386
		MCP	0.000	0.759	0.241	3.508	0.492	0.4629244	0.4616298
		LASSO	0.000	0.063	0.937	1.642	2.358	0.4617325	0.4592168
		LAD-LASSO	0.000	0.136	0.864	1.857	2.143	0.4931865	0.4895968
		ORACLE	0.000	1.000	0.000	4.000	0.000	0.4955971	0.4927684
0.5	200	SCAD	0.000	0.775	0.225	3.519	0.481	0.4694942	0.4690124
		MCP	0.000	0.800	0.200	3.592	0.408	0.4695161	0.4692473
		LASSO	0.000	0.065	0.935	1.632	2.368	0.4689443	0.4684985
		LAD-LASSO	0.000	0.096	0.904	1.806	2.194	0.4840470	0.4825058
		ORACLE	0.000	1.000	0.000	4.000	0.000	0.4851959	0.4843832

least overfitted model, also has the most correct number of zeros. Here, LASSO has the lowest average MAPE, SCAD has the lowest median MAPE. By looking at the results, we see that LAD-LASSO has the highest average and median MAPE, for n =200 this trend

Table 3: Simulation results for t3 error

σ	n	**Method**	**Underfitted**	**Correctly Fitted**	**Overfitted**	**Correct_0s**	**Incorrect_0s**	**Avg_MAPE**	**Median_MAPE**
1	50	SCAD	0.211	0.199	0.590	2.565	1.435	1.076633	1.045244
		MCP	0.281	0.270	0.449	2.906	1.094	1.078710	1.047564
		LASSO	0.121	0.059	0.820	1.690	2.310	1.072584	1.039920
		LAD-LASSO	0.134	0.104	0.762	1.933	2.067	1.118372	1.091138
		ORACLE	0.070	0.930	0.000	4.000	0.000	1.120909	1.099607
1	100	SCAD	0.076	0.278	0.646	2.628	1.372	1.100795	1.074740
		MCP	0.127	0.416	0.457	3.039	0.961	1.102760	1.077400
		LASSO	0.026	0.070	0.904	1.651	2.349	1.099866	1.077130
		LAD-LASSO	0.027	0.086	0.887	1.684	2.316	1.118307	1.096043
		ORACLE	0.014	0.986	0.000	4.000	0.000	1.117215	1.099615
1	200	SCAD	0.026	0.329	0.645	2.731	1.269	1.106202	1.096330
		MCP	0.036	0.524	0.440	3.102	0.898	1.106711	1.095518
		LASSO	0.005	0.086	0.909	1.633	2.367	1.107248	1.098531
		LAD-LASSO	0.005	0.085	0.910	1.658	2.342	1.116069	1.105823
		ORACLE	0.002	0.998	0.000	4.000	0.000	1.114268	1.107052
0.5	50	SCAD	0.025	0.396	0.579	2.945	1.055	0.5350297	0.5183675
		MCP	0.039	0.582	0.379	3.255	0.745	0.5356262	0.5194496
		LASSO	0.006	0.097	0.897	1.647	2.353	0.5337229	0.5168493
		LAD-LASSO	0.010	0.160	0.830	1.942	2.058	0.5975175	0.5830912
		ORACLE	0.004	0.996	0.000	4.000	0.000	0.6007809	0.5899422
0.5	100	SCAD	0.005	0.554	0.441	3.226	0.774	0.5428540	0.5331059
		MCP	0.008	0.698	0.294	3.434	0.566	0.5435128	0.5332876
		LASSO	0.003	0.067	0.930	1.591	2.409	0.5438373	0.5349139
		LAD-LASSO	0.003	0.105	0.892	1.805	2.195	0.5733441	0.5643153
		ORACLE	0.001	0.999	0.000	4.000	0.000	0.5734474	0.5663596
0.5	200	SCAD	0.000	0.720	0.280	3.477	0.523	0.5509055	0.5464532
		MCP	0.001	0.786	0.213	3.575	0.425	0.5510842	0.5463737
		LASSO	0.000	0.077	0.923	1.690	2.310	0.5522502	0.5467197
		LAD-LASSO	0.000	0.118	0.882	1.798	2.202	0.5663444	0.5608860
		ORACLE	0.000	1.000	0.000	4.000	0.000	0.5652286	0.5613562

follows. We can say in presence of double exponential error with high noise LAD-LASSO performs slightly better than LASSO, although SCAD and MCP model performs well, even better than LAD-Lasso.

In the same table for low noise ratio, it is found that LAD-LASSO has higher correctly fitted models than LASSO across all the sample size. MCP performs very well, has the highest percentage of correctly fitted model among all the different settings in table1, LAD-lasso has more number of correctly zeros identified than LASSO, in these four model, lasso has the most number of incorrect zeros. MCP and SCAD has the similar and least average and median MAPE.

Across all scenarios, ORACLE consistently achieves the best performance, with the highest correctly fitted values and minimal incorrect zeros and overfitting, demonstrating its theoretical property. SCAD and MCP generally perform well, with MCP showing a slight edge in terms of higher correctly fitted values and fewer incorrect zeros at larger sample sizes. LASSO tends to overfit significantly across all sample sizes, as reflected by its consistently high overfitting rates, though it achieves comparable average and median MAPE values to other methods. LAD-LASSO shows balanced performance with moderate overfitting and stable average MAPE, particularly as n increases. Notably, as sample size increases, all methods improve in terms of correctly fitted values and reduced underfitting, reflecting the benefit of larger datasets for accuracy. Lower noise levels (0.5) yield significantly lower MAPE values compared to higher noise levels (1), stating the effect of noise on model accuracy.

In table 2, with heavy tailed error t distribution with 5 df with high ratio of noise level, we observe that performance of the model increases with larger sample sizes, as we can see smallest sample size with high noise ration yields more underfitted models, similarly the proportion of correctly fitted model is significantly higher across with larger

sample size and low level of noise ratio. These findings emphasize the strengths of MCP and SCAD for practical applications, particularly in low-noise and larger sample size scenarios, while also demonstrating the limitations of LASSO in terms of overfitting and the robust yet consistent performance of LAD-LASSO. Once again we see that MAPE's are significantly lower than in low noise level.

From table 3 2, with heavy tailed error t distribution with 3 df we observe that for n=50, sigma=1 LASSO has the least proportion of Underfitted, MCP has the largest proportion of correctly fitted and least overfitted model, also has the most correct number of zeros. Here, LASSO has the lowest average MAPE, SCAD has the lowest median MAPE as well. By looking at the results, we see that LAD-LASSO has the highest average and median MAPE here as well. For n=100 LASSO has the least proportion of Underfitted, MCP has the largest proportion of correctly fitted and least overfitted model, also has the most correct number of zeros. Here, LASSO has the lowest average MAPE, SCAD has the lowest median MAPE. By looking at the results, we see that LAD-LASSO has the highest average and median MAPE, for n =200 this trend follows. LAD-LASSO consistently achieves a higher percentage of correctly fitted models compared to LASSO across all sample sizes. MCP demonstrates good performance, achieving the highest percentage of correctly fitted models across all settings. Additionally, LAD-LASSO identifies more correct zeros than LASSO, while LASSO shows the highest number of incorrect zeros among the four models. Both MCP and SCAD show similar performance, achieving the lowest average and median MAPE values.

4 Conclusion

In the original paper by (Wang, Li, and Jian) AIC, AICs, and BIC method was compared with Lad-lasso to assess the performance. Given the course material and my understanding, I introduced SCAD, MCP and Lasso to assess the model performance. Even with the heavy tailed error distribution SCAD and MCP perform better than LAD-Lasso in accuracy, however LAD-LASSO offers a good balance and works well across different situations. This makes it a practical and flexible choice for regression analysis. LAD-LASSO consistently finds more correctly fitted models and identifies correct zeros better than LASSO, also the effect of noise is notable. Future studies could explore ways to improve LAD-LASSO further or test it on more complex data.

References

Hansheng Wang, Guodong Li, and Guohua Jiang (2007) "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso" ,Journal of Business and Economic Statistics, July 2007, pp. 347-355

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in 2nd International Symposium on Information The ory, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267-281. Bassett, G., and Koenker, R. (1978), 'Asymptotic Theory of Least Absolute Error Regression," Journal of the American Statistical Association, 73, 618-621.

Bloomfield, P., and Steiger, W. L. (1983), Least Absolute Deviation: Theory, Applications and Algorithms, Boston: Birkhauser. Craven, P., and Wahba, G. (1979), "Smoothing Noise Data With Spline Func tion: Estimating the Correct Degree of Smoothing by the Method of Gener alized Cross Validation," Numerische Mathematik, 31, 337-403.

Davis, R. A.. Knight, K., and Liu, J. (1992), "M-Estimation for Autoregres sions With Infinite Variance," Stochastic Process and Their Applications, 40, 145-180.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Like lihood and Its Oracle Properties," Journal of the American Statistical Asso ciation,96, 1348-1360.

Hurvich, C. M., and Tsai, C. L. (1989), "Regression and Time Series Model Selection in Small Samples," Biometrika, 76, 297-307.

- (1990), "Model Selection for Least Absolute Deviation Regression in Small Samples," Statistics and Probability Letters, 9, 259-265. Knight, K. (1998), "Limiting Distributions for L Regression Estimators Under General Conditions," The Annals of Statistics, 26, 755-770. Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," The Annals of Statistics, 28, 1356-1378.

Koenker, R., and Zhao, Q. (1996), "Conditional Quantile Estimation and Infer ence for ARCH Models," Econometric Theory, 12, 793-813.

Ling, S. (2005), "Self-Weighted Least Absolute Deviation Estimation for Infi nite Variance Autoregressive Models," Journal of the Royal Statistical Soci ety, Ser. B, 67, 1-13.

McQuarrie, D. R., and Tsai, C. L. (1998), Regression and Time Series Model Selection, Singapore: World Scientific.

Nissim, D., and Penman, S. (2001), "Ratio Analysis and Equity Valuation: From Research to Practice," Review of Accounting Studies, 6, 109-154. Peng, L., and Yao, Q. (2003), "Least Absolute Deviation Estimation for ARCH and GARCH Models," Biometrika, 90, 967-975.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Es timators," Econometric Theory, 7, 186-199.

Schwarz, G. (1978), "Estimating the Dimension of a Model," The Annals of Statistics, 6, 461-464.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," Statistica Sinica, 7, 221-264.

Shi, P., and Tsai, C. L. (2002), "Regression Model Selection: A Residual Likelihood Approach," Journal of the Royal Statistical Society, Ser. B, 64, 237-252.

- (2004), "A Joint Regression Variable and Autoregressive Order Selec tion Criterion," Journal of Time Series, 25, 923-941. Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," Journal of the Royal Statistical Society, Ser. B, 58, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Spar sity and Smoothness via the Fused Lasso," Journal of the Royal Statistical Society, Ser. B, 67, 91-108.

Zou, H., Hastie, T., and Tibshirani, R. (2004), "On the 'Degrees of Freedom' of Lasso," technical report, Standford University, Statistics Department.